融合词语义表示和新词发现的领域本体演化*
——以产品评论数据为例

■ 耿骞[1,2] 邓斯予[2] 靳健[2]
[1] 北京师范大学政府治理研究中心 珠海 519087 [2]北京师范大学政府管理学院 北京 100875

摘 要:[目的/意义]针对传统本体演化中对新知识和新需求的捕捉存在不准确、低效率的问题,提出一种基于领域新词发现的本体演化方法,并以用户产品评论数据为例进行验证。[方法/过程]首先采用自然语言处理算法对用户产品评论文本语料进行文本预处理,并利用 Word2vec 算法进行词向量嵌入;然后采用深度学习中 Bi-LSTM-Attention-CRF 算法实现候选领域新词的识别和抽取,并利用 K-means 算法进行聚类以得到最终领域新词;最后利用本体演化的六阶段演化流程,实现领域本体的演化工作。[结果/结论]以智能手机领域产品评论为实验数据,验证了本研究采用领域新词发现模型具有更高的准确率和召回率,由此演化得到智能手机领域新版产品本体。领域新版产品本体既可以帮助产品设计者根据领域本体中新特征、新功能优化产品设计,也可以支持消费者利用产品评论进行购买决策。

关键词:本体演化 领域新词 新词发现 注意力机制 双向长短期记忆网络 条件随机场
分类号:G250 TP391.1
DOI:10.13266/j.issn.0252-3116.2021.08.009

1 引言

随着亚马逊、淘宝等电商网站的快速发展,用户可以更容易表达对于各类产品的评论。这些用户评论可以帮助潜在消费者获取到有关产品某些特征的情感倾向,以支持他们的购买决策;商家和产品设计师也可以根据这些产品评论改善服务,提升产品质量[1]。然而,用户产品评论是非结构化、内容规模庞大的文本数据[2],蕴含着多种多样的实体及实体间复杂的隐含关系[3]。本体(Ontology)作为知识组织的工具,在众多实际应用中开发并广泛使用。利用领域本体,可以实现对用户产品评论的知识组织、知识存储以及知识应用,为深入挖掘产品评论内容提供支持。

现实世界中知识在不断地更新,用户对知识的需求也处于不断变化的过程中。随着产品发布新功能、新特征,用户评论也会随之发生改变。例如,苹果公司在2019年推出的 iPhone 11 系列手机,在手机的摄像头上出现新的特征:“浴霸”摄像头;2017年发行的 iPhone X 也出现新的面部识别的新功能。这些手机产品的新特征和新功能都是用户评论中会集中关注的热门话题,而用户评论中出现的实体词和特征词可能在已有的领域产品本体中并不存在,此时需要对原版本的领域本体进行演化以满足新需求。本体演化(Ontology Evolution),也叫本体进化,正是修改现有本体以适应新知识和变化的需求[4]。这种演化体现了本体的持久性,利于在实际应用中长期发挥其价值。并且,本体规模庞大且不断发展,导致本体演化是一项复杂且耗时的任务[5-6]。

在本体演化中,变化的捕捉是整个流程中核心的步骤,如果可以利用前沿的算法实现从新的评论文本语料中,自动化识别并抽取出领域新词,并将这些领域新词用作本体演化需要捕捉的变化,对于从评论文本中构建领域产品本体的演化而言具有重大的意义:既可以使产品设计者实时掌握产品评论中消费者关注的

* 本文系国家社会科学基金重点项目“面向集成管理的政府数据组织与传递机制研究”(项目编号:19ATQ005)和国家自然科学基金项目“差异化客户需求的提取及比较研究:基于产品在线评论的挖掘分析”(项目编号:71701019)研究成果之一。
作者简介:耿骞(ORCID:0000-0001-5064-4996),教授,博士,博士生导师;邓斯予(ORCID:0000-0002-8000-4065),硕士研究生;靳健(ORCID:0000-0002-3239-2294),副教授,博士,硕士生导师,通讯作者,E-mail:jinjian.jay@bnu.edu.cn。
收稿日期:2020-10-15 修回日期:2021-01-12 本文起止页码:85-96 本文责任编辑:杜杏叶

热门功能、组件，以优化产品设计；也可以为消费者在购买产品时，利用新颖的用户产品评论做出购买决策提供支持。因此，本研究在领域本体演化中引入领域新词发现技术。领域新词发现异于传统新词发现，要发现的新词可能只是在某一个领域中从未出现，而并非所有领域；发现领域新词可以挖掘出该领域最新的发展动态。例如对某类产品的用户评论中领域新词发现，可以帮助人们了解该类产品当前最新出现的功能、成分、包装等。现如今，神经网络中的深度学习技术备受关注且发展迅速。深度学习方法用于学习样本数据的内在规则和表示水平，并发现数据中的隐藏模式[7]；并且，当数据集大小增加时，基于深度学习的方法往往表现更好[8]，例如在自然语言处理中的分词、命名实体识别等研究中。因此，为了实现领域新词发现，本文利用深度学习的算法模型处理新文本语料中的数据，并对新特征进行自动化识别和提取，进而对构建的本体进行结构和内容的调整，以支持本体演化中变化捕捉工作，实现对于产品设计者和消费者的帮助和支持。

2 相关研究现状

2.1 本体演化和新词发现研究现状

在国内外已有研究中，在本体演化和新词发现都分别有一些研究进展，但两者结合的研究却相对较少。

首先，本体演化是在保证本体一致性的前提下，对本体所作的一系列修改过程，它可以被看作为本体发展过程中一系列操作的结果。在本体演化的研究中，可以总体归纳为手动化演化方法、半自动化演化方法、自动化演化模型或系统这三类。在手动化演化研究中，当新知识或者新的需求产生时，V. S. K. Nagireddi 等[9]和 X. Chen 等[10]利用领域专家进行演变，或者将已有本体与其他领域本体进行合并。在半自动化演化研究中，刘紫玉等[11]提出了基于 DBpedia 的本体半自动化演化方法；陈晶等[12]基于邻接表的 SPFA 算法优化波及效应的计算，并使用 Floyd-Warshall 算法对规模较大的本体进行评估。在自动化演化研究中，刘毅等[13]提出一个本体演化驱动的语义搜索引擎系统——OESSE，将本体自动进化功能与语义搜索进行了有机结合；刘莹[14]将知识管理的社会性融入到应用技术之中，提出了一个基于知识本体演化和信息检索联动发展的分布式知识管理系统；C. Huang 等[15]针对智能制造应用程序实现的目标，提出了一种本体生成和演化的系统，该系统可以自动从原始生产数据中提取本体，并根据制造数据环境的变化动态调整本体。

其次，领域新词发现就是在某一特定领域内，之前从未出现过的新词的识别和抽取的过程。在传统的新词发现研究中，互信息和邻接熵被引入新词发现的研究中[16-18]。杜丽萍[19]提出了基于 PMI^k 算法与少量基本规则的互信息改进算法，验证了通过进行新词发现能有效改善分词系统对网络文本的处理效果。也有学者基于规则的方法进行新词发现，如周霜霜[20]融合规则和统计的方法进行微博新词发现，王馨[21]采用了关联规则对网络新闻热点进行排名，陈梅婕[22]在利用双向聚合度时采用了词边界筛选规则，进而提升了专利新词发现性能。此外，还有学者在新词发现研究中引入了前沿的算法和模型。其中，张华平等[23]采用条件随机场对社会媒体领域的大规模语料中的新词进行预测，取得了更快的速度及更高的精度；王汀等[24]融合条件随机场和支持向量机的方法进行新词发现，在获取中文百科分类页面中的实体识别时取得更高的查准率和查全率；陈先来[25]在利用新词发现改进现有分词模型时，采用了互信息和逻辑回归算法，提高医学文本分词的准确率；刘昱彤[26]在从大规模古汉语语料中发现新词的研究中改进了 Apriori 算法，并加入了长短期记忆网络和条件随机场算法，该方法经验证可以在宋词和宋史数据集上有限的识别出新词；赵志滨[27]将句法分析和词向量用于新词发现的研究，通过护肤品论坛的真实文本数据集验证了该方法进行新词发现具有良好的性能；黄文明[28]采用信息量和 Bi-LSTM + CRF 算法进行领域新词发现，通过联想客服问答系统的问题数据集验证了该方法可以提高了领域新词识别的准确率。

综上所述，国内外已有一些关于本体演化和新词发现的研究。然而，已有的研究对于大数据规模的本体预料进行演化处理时，基本没有结合新词发现的研究方法，其演化效果往往表现较差；其次，在领域新词发现的研究中，大多数没有结合深度学习中的一些前沿算法和模型，在处理数据量庞大、关系复杂的数据预料时，往往很难精准而快速的发现领域新词。此外，在对于非结构化文本数据的知识组织的研究中，构建出的本体往往不具有长久性和实时性，随着时间的推移以及开发者后期维护较少，基本难以得以发挥持续价值。因此，本研究将通过基于领域新词发现的本体演化技术，实现对用户产品评论中构建的领域本体的演化工作，以充分发挥领域本体的应用价值，为消费者利用产品评论进行购买决策提供支持。

2.2 关键技术现状

2.2.1 LSTM 网络

长短期记忆网络(Long Short-Term Memory,简称 LSTM)是循环神经网络(Recurrent Neural Network,简称 RNN)的一种,具有记忆数据序列的能力。RNN[29] 主要由输入层、隐藏层和输出层构成,具有记忆当前输入和上文输入的信息的功能,并且在处理短时间序列的文本序列时表现更好。然而,在处理长时间序列信息时,RNN 可能会出现梯度消失或爆炸的问题。因此,A. Graves[30] 提出的 LSTM 神经网络解决了 RNN 存在的问题,并在图像处理和语音识别等领域广泛使用。

相比于 RNN,LSTM 在其结构基础上增加了记忆单元,以及由输入门、遗忘门和输出门构成的三种控制门结构[31],见图 1。门结构是神经网络中的一层全连接层,输入向量由门结构处理后输出 0 到 1 之间的实数向量。LSTM 这种门结构基于 sigmoid 函数,从而使得神经网络拥有允许数据通过(选择性保留)或丢弃状态值的功能,便于获取长期序列距离中的文本序列。

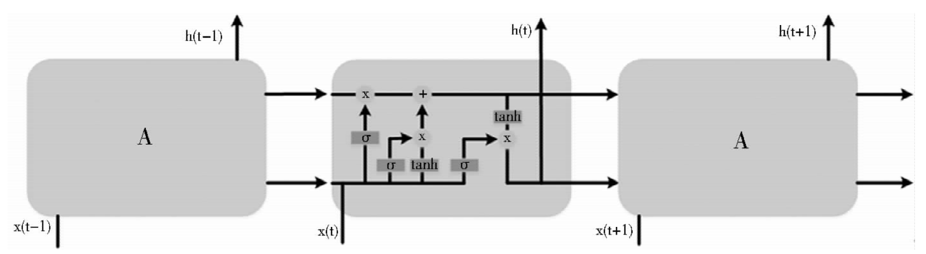


图 1 LSTM 的结构示意

此外,在文本序列预测的时候,有的预测结果可能由前面若干输入和后面若干输入共同决定,由此出现了类似于双向循环神经网络(BRNN)的双向长短期记忆网络(简称 Bi-LSTM)。Bi-LSTM 主要包括前向传播和后向传播两个过程:将训练序列输入到前向 LSTM 网络模型,通过前向传播计算得到前向特征信息;同样地,输入后向 LSTM 网络模型,通过后向传播计算得到后向特征信息,再将前向特征信息与后向特征信息拼接获得最终的隐藏状态,这样就汇总了前向和后向双向语义特征。利用 Bi-LSTM 在解决关系复杂的文本序列时,例如在用户产品评论中,由于文本序列的预测不仅取决于序列前面一些输入文本,同时也会受到序列后面输入文本的影响,所以可以采用 Bi-LSTM 提高评论文本的预测准确率。

2.2.2 Attention 机制

传统编码－解码器模型(Encoder-Decoder Model)主要用于处理文字、语音、图像、视频等数据,由此衍生出 RNN、LSTM 等算法。在处理文本序列时,编码器将输入文本序列编码成固定长度的隐向量,并对隐向量赋予相同的权重;解码器基于这些隐向量解码输出。当输入序列文本内容扩大,且文本序列对应的分量权重相同时,Encoder-Decoder 模型对于输入文本序列的区分度下降,造成模型性能也随之下降。因此,D. Bahdanau 等[32] 提出了 Attention 注意力机制可以很好地解决此缺陷。Attention 机制用于提升编码－解码器模型效果,从大量信息中快速筛选出高价值信息,其本质是模拟人的注意力,仿照着人类在观察物体时大脑的思维活动[33]。因此,Attention 机制在情感分类、机器翻译等多个研究领域都有重要的应用价值。

在编码－解码器模型的优化中,Attention 机制主要用于解码过程,它改变了传统 Decoder 对每一个输入输入文本序列都赋予相同向量的缺点,而是根据单词的不同赋予不同的权重。在 Encoder 过程中,输出不再是一个固定长度的中间语义,而是一个由不同长度向量构成的文本序列,Attention 机制使得模型对输入文本序列的不同时刻隐向量赋予了相对应的权重,并按重要程度将隐向量合并为新的隐向量,最后输入到 Decoder 中;而 Decoder 过程根据这个序列子集进行进一步筛选和处理。因此,引入 Attention 机制的 Encoder-Decoder 模型如图 2 所示:

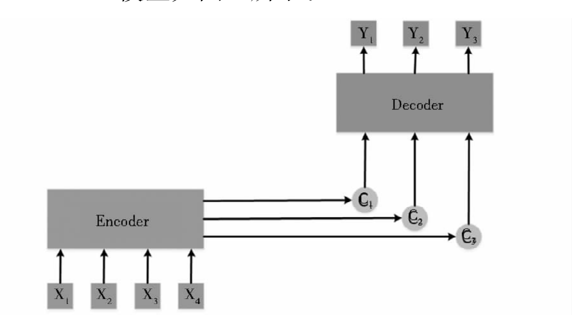


图 2 引入 Attention 机制后的 Encoder-Decoder 模型

2.2.3 CRF 序列标注

条件随机场(Conditional Random Field,简称 CRF),结合了最大熵模型和隐马尔可夫模型的特点,

是一种无向图模型。CRF 模型在给定一组输入随机变量 X 条件下,给出另外一组输出随机变量 Y 的条件概率分布模型,并且已被应用于序列标注的不同预测任务上[34]。CRF 模型基本流程如图 3 所示:

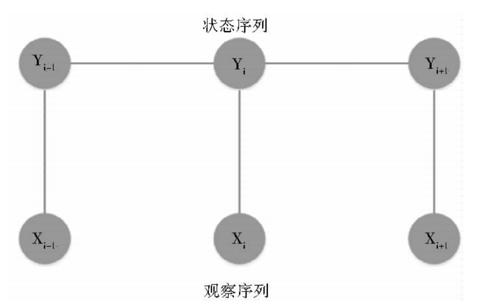


图 3 CRF 模型示意

CRF 序列标注在实现前,需要由人工标注原始语料信息,人为地定义预料中词的词性、程度、类别等属性。现阶段,在进行一些自然语言处理时,例如命名实体识别工作,会采用神经网络模型学习训练数据,并产生特征向量,以获得更好的预测效果。然而神经网络模型会比较耗时,且模型的部分输出结果是错误的识别结果。所以 CRF 模型可以用于命名实体识别任务中,将一些人工预定义的规则添加到序列标记过程中,这样可以取得更好的预测效果。

综上所述,LSTM 记忆单元和门结构有效解决了传统 RNN 中的梯度消失缺陷;双向 LSTM 模型不仅能识别过去的文本序列信息,还能充分考虑未来的序列信息,使得上下文信息充分完整被利用。在编码 – 解码模型中引入 Attention 机制,很好解决了文本序列的长度扩张时,各个序列部分的权重。CRF 序列标注关注了整个文本序列的局部特征的线性加权组合,即通过特征模板扫描整个句子,计算的是联合概率,优化了整个序列。所以,在进行大数据量的用户评论文本处理时,可以引入 Bi-LSTM 神经网络、Attention 机制和 CRF 序列标注,实现更准确的文本实体识别效果。

3 基于新词发现的本体演化

本研究提出基于领域新词发现的本体演化框架,见图 4。其核心在于在本体演化中,加入领域新词发现,用于捕捉本体的变化。而领域新词发现主要采用深度学习中基于注意力机制的双向长短期记忆神经网络结合条件随机场模型(Bi-LSTM + Attention + CRF)。

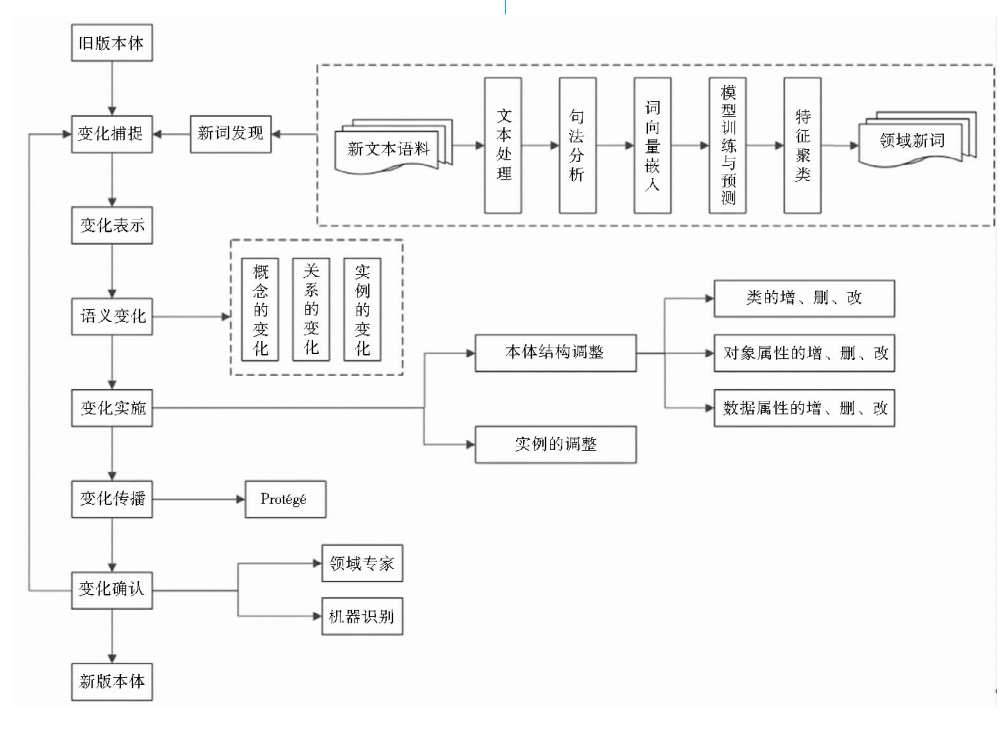


图 4 基于新词发现的本体演化整体框架

3.1 基于深度学习的新词发现

如上文所述,已有众多研究提出了不同的新词发现的方法,如融合信息量[28]、互信息[18-19]、句法分析[27]、规则[22, 24]等方法。本研究为进一步提升新词发现的准确率,主要采用深度学习中一些前沿算法模型,如 Word2vec 算法、Bi-LSTM-Attention-CRF 模型等。根据图 4 中基本流程,基于深度学习的新词发现主要包含文本预处理、句法分析、词向量嵌入、模型的训练与预测、特征聚类等五个步骤。由此本研究提出一个用于本体演化的新词发现新框架,如图 5 所示:

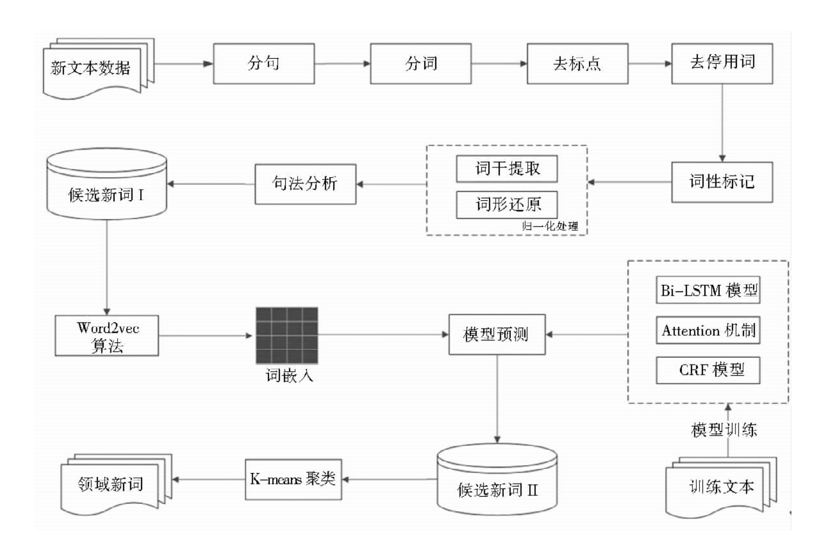


图 5　用于本体演化的新词发现流程

3.1.1　文本预处理

本研究以亚马逊电商平台(Amazon. com)的用户产品评论数据为研究数据。在进行新词发现前,需要对原始英文预料进行文本预处理,以去除原始文本预料中的异常字符和标点等。本研究以 Python 为开发语言,采用 Python 中的 NLTK 工具库进行分句(Sentence Segmentation)、分词(Word Tokenization)、去标点(Eliminating Punctuation Marks)、去停用词(Removing Stop Words)、词性标记(POS tagging)等,并在词性标记的基础上,对句子中的词进行归一化处理(Normalization)——词干提取(Stemming)和词性还原(Lemmatization)。最后得到带有词性标注的原始单词。

本研究选用了两组数据集,分别作为训练集和测试集。训练集主要用于训练 Bi-LSTM-Attention-CRF 模型,测试集主要来源于新文本语料,用以进行领域新词的发现和抽取。

3.1.2　句法分析

在对原始预料进行预处理的基础上,引入句法分析可以实现对领域候选新词的第一次筛选和抽取。句法分析(Syntactic Parsing),作为自然语言处理中关键低层技术之一,是对句子中的词语语法功能进行分析[3]。句法分析分为句法结构分析(Syntactic Structure Parsing)和依存关系分析(Dependency Parsing)。为获取整个句子的句法结构或者完全短语结构为目的的句法分析,被称为句法结构分析;而以获取局部成分为目的的句法分析,被称为依存分析。

在用户产品评论文本中,领域新词主要是由名词、动词构成的特征词,以及特征词之间的关系词。因此,为获取文本句子中这部分成分,本研究采用依存分析。依存分析通过分析语言单位内成分之间的依存关系揭示其句法结构。直观来讲,依存句法分析识别句子中的"主谓宾""定状补"这些语法成分,并分析各成分之间的关系。比如"Wireless charging damages battery health",这里"Wireless charging"是主语,"damages"是谓语,"battery health"是宾语,这里的主语和宾语都有可能成为领域新词,因此可以作为候选领域新词。因此,通过句法分析可以得到在文本句子中可能会成为领域新词的候选集,并人工定义筛选规则进行过滤,筛选出第一阶段的候选新词。

3.1.3　词嵌入

在深度学习中,利用词嵌入(Word Embedding)的特征学习是抽取实体的有效方式[28]。在利用深度学习模型进行数据训练和预测前,需要首先进行的工作就是词嵌入,将经过文本预处理后的词转为数值化向量的过程,即词向量化。

Word2vec(Word to Vector)是一个开源的深度学习工具,用于基于神经网络语言模型和对数双线性模型计算单词向量[35]。通过学习文本,捕获文本中单词的语义信息,并用词向量的方式表示单词。Word2Vec 主要包括两个模型:CBOW 和 Skip-Gram。CBOW 模型根据给定的上下文预测目标单词信息,而 Skip-Gram 模型则根据给定的单词预测在其上下文中出现的单词。

因此,考虑到用户产品评论中可能具有多种复杂且潜在的特征和特征间关系,例如用户评论数据"It's not cool for me that AppStore occupy much storage, and waste battery quickly"中,特征词"AppStore""storage"和"battery"三者之间是存在着相互交错的关联关系,且在上下文中可以获取到对于一个特征词的关联关

系,因此,为利于提升领域新词发现的精准度,本研究采用 Word2vec 中的 CBOW 模型进行词嵌入,并在降低训练复杂度时采用负杂样(Negative Sampling),以实现通过训练评论文本的上下文语境,预测出某一个词汇的词向量表示。

3.1.4 Bi-LSTM-Attention-CRF 模型训练

在词嵌入的基础上,考虑到本研究的研究数据是大规模数据量的非结构的评论文本,其中存在复杂多样的特征实体和潜在的关联关系,用传统的命名实体识别的方法在效率和效果层面都存在不足。此外,深度学习算法,例如 Bi-LSTM + CRF 算法,在获取领域新词时效果往往更加准确[28]。因此本研究引入 Bi-LSTM-Attention-CRF 模型,即基于注意力机制的双向长短期记忆神经网络结合条件随机场模型。本研究中引入的 Bi-LSTM-Attention-CRF 模型框架如图 6 所示:

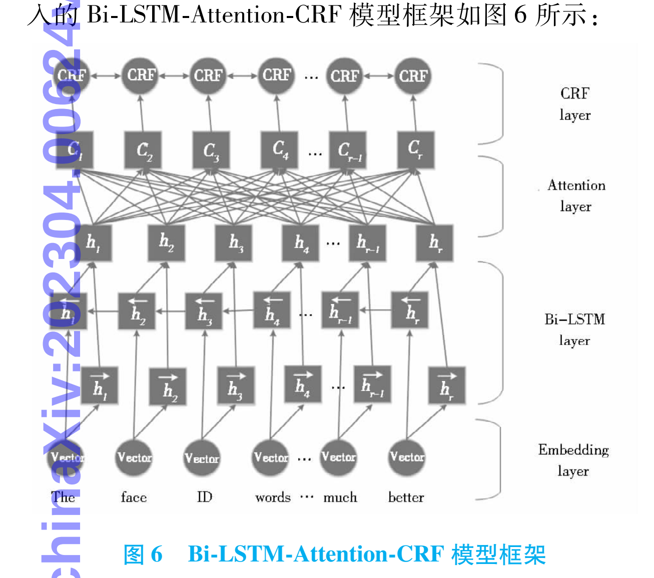


图 6 Bi-LSTM-Attention-CRF 模型框架

Bi-LSTM-Attention-CRF 模型具体实现是,通过保存双向 LSTM 中 Encoder 编码器对输入文本序列的中间输出,训练一个模型来选择性学习这些文本输入,并在模型输出时将输出序列与之进行关联;然后双向 LSTM 加上 Attention 机制学习到了输入文本序列中前后信息的特征,再利用 CRF 模型根据给定观察序列推测对应的状态序列,可以利用相邻前后的标签关系来获取当前的最优的标记。

在本研究中,为发现输入文本中的领域新词,采用 Bi-LSTM-Attention-CRF 模型来处理文本预处理后的训练集和测试集。首先,输入文本序列中每个词的词向量,以及自动标注的数据集;其次对训练集中领域词向量进行模型训练;然后对经过分词处理后的原始训练集,进行人工标注领域新词,其采用的标注标签为 BIESO 五种标签(B_new, I_new, E_new, S_new, O)。其中,B 即 Begin,表示新词词组的开始;I,即 Intermediate,表示新词词组的中间;E,即 End,表示新词词组的结尾;S,即 Single,表示单个新词字符;O,即 Other,表示其他,用于标记非新词的无关字符。数据集进行标注的示例如表 1 所示:

表 1 标注示例

词	标签
3D	B_new
Touch	E_new
and	O
face	B_new
recognition	E_new
are	O
super	O
fast	O
and	O
user	O
friendly	O

最后,利用训练集训练 Bi-LSTM-CRF 网络,即每一轮迭代都需要进行 Bi-LSTM 前向传播和后向传播、Attention 层编码和解码、CRF 层正向传播和反向传播,并用训练后的模型预测测试集中的领域新词。通过 Bi-LSTM-Attention-CRF 模型对测试集数据进行领域新词的预测后,获取到第二阶段的领域候选新词。

3.1.5 特征聚类

由于用户产品评论中,对于产品同一个特征、功能或组件的表达方式不尽相同,例如 iPhone X 系列出现的新功能——面部识别,有的用户可能会用"face recognization""facial recognization""face scanning""facial scan"等词组。此时需要对第二阶段识别出的候选领域新词进行过滤,筛选出真正的用于本体演化的领域新特征,本文则采用同义词中高频词作为该特征的领域新词。此外,对于第二阶段获取到的领域候选新词,需要对这些新词进行归类,判断其隶属于产品本体中哪个位置,利于后期本体演化工作的变化捕捉,本文则采用特征抽取中 K-means 聚类方法。

K-Means 算法的基本思想是,事先确定常数 K,常数 K 意味着最终的聚类类别数。首先随机选定初始点为质心,并通过计算每一个样本与质心之间的相似度,将样本点归到最相似的类中;然后重新计算每个类的质心,重复这样的过程,直到质心不再改变,最终就确定了每个样本所属的类别以及每个类的质心。通过 K-means 可以将候选领域新词进行初步聚类,然后通过领域专家知识进行类别的判断,以确定出一个产品

不同类别下领域新特征。如表 2 所示为 iPhone 手机产品的新特征聚类结果示例。

表 2　新特征聚类结果示例

类别	特征词			
feature	iPhone X	iPhone Xr	256G memory	iPhone 11
function	3D Touch	face recognition	face ID	battery health
component	crystal clear speaker	glass screen protector	AirPods	no-home button

3.2　基于新词发现的本体演化

根据图 4 基本流程图可知，本研究选用的本体演化基本框架，即从变化捕捉(Capturing)、变化表示(Representation)、语义变化(Semantic of change)、变化实施(Implementation)、变化传播(Propagation)到变化确认(Validation)这一基本流程框架，主要采用的是在本体演化流程的研究中具有代表性的 L. Stojanovic 等[37]提出的六阶段划分法。

本研究提出的本体演化的框架在六阶段划分法的基础上，首先在变化捕捉过程中增加了新词发现技术，如 3.1 小节所述。由于本研究的对象是评论文本，且文本数据随着时间和空间的变化，会出现较多某一个领域的新特征、新功能。在用户产品评论中会包含关于新款产品的最新功能和组件，例如 iPhone X 出现的无线充电、无 Home 按键等。因此利用新词发现技术，可以很好地捕捉本体中概念和关系以及实例的变化。

其次，变化表示(Representation)是处理变化的前导工作，实质是用形式化的方式表示领域本体的变化动作，包括领域本体结构、概念的调整，例如利用产品评论中一些典型的领域特征词表示产品某一方面的特征、功能等。语义变化(Semantics of change)是对领域本体变化进行语义控制，包括概念的变化、关系的变化以及实例的变化。在领域产品本体中，概念的变化主要体现在对于领域本体的类的调整，例如手机插口类(Jack class)在原领域本体中是包含耳机插口和充电插口两个子类(Subclass)，而新产品(如 iPhone 8 系列、iPhone X 系列)中将充电插口和耳机插口合并，共用一个插口，此时需要调整该概念为充电及耳机插口。关系的变化主要体现在，原本体中存在的一对一的关系会被调整为一对多、多对一，甚至多对多的关系。例如新产品的价格类(Price class)会出现多个类共同决定价格，即产地(place of origin)、内存(storage)、颜色(color)、屏幕尺寸(screen size)等会决定该手机产品的价格。实例的变化主要是出现的一些新实例(individual)，例如手机像素类(camera_pixel)出现单摄 1 200 万像素、后置双 1 200 万像素等；此外，iPhone 11 系列产品的机身颜色(body_color)出现了紫色、白色、绿色、黄色、黑色、红色六种颜色实例。

变化实施(Implementation)的工作包括对于本体结构的调整和实例的调整。其中，对于本体结构的调整，是包括类的增删改、对象属性的增删改、数据属性的增删改。对于实例的调整，本研究主要采用 C. Huang 等[15]提出对于实例增加和调整约束(Restriction)的方法。对于领域产品本体的调整如表 3 所示。变化的传播(Propagation)是在一个领域本体发生演化后保证并维护和它相关的本体的一致性，以避免本体演化造成的重要影响之一——导致前后本体版本的不兼容。本研究采用 protégé 中的演化插件，如 Change-management 插件、PROMPT 插件[38]等，对这些变化进行传播和转移，以便于其他领域本体的重用和继承本领域的新版本体。

表 3　领域产品本体中类和实例的调整(部分)示例

	类 Class		实例 Individuals	
增加	face_recognition	Wireless_charging	AirPower	256G
删除	fingerprint_recognition	Home_button	4_inches	16G
修改	Metal_frame→Stainless_steel_frame	Retina_Display→Super_Retina_Display	A6_processor→A11_processor	8_million_pixels→12_million_pixels

最后，变化的确认(Validation)阶段是对上述领域本体演化过程的最终确认，通过领域专家或机器识别的方法，对以上这些步骤的核准之后，确认对领域本体的修改，并且还可根据从文本中挖掘的用户需求删除一些变化，以完成变化的最终确认。

4　实验与结果分析

4.1　数据来源与预处理

本文以智能手机领域用户产品评论对例进行实验研究。在前期研究[2-3]已构建出智能手机领域产品本体的基础上，本实验选用了亚马逊电商平台(Amazon.com)的苹果公司 iPhone 智能手机 2019 年新款产品评

论为研究数据,对前期研究构建的领域本体进行本体演化的实验。利用爬虫共爬取 2013 年款 iPhone 5C/5S 系列手机评论共 10 437 条,和 2019 年新款 iPhone11 系列评论共 2 798 条,作为本研究的实验数据。其中以 2013 年款 iPhone 5C/5S 系列手机评论数据为训练集数据,以 2019 年新款 iPhone11 系列评论为测试集数据。

利用 3.1.1 小节的文本预处理方法,对新语料库进行分句、分词、去标点、去停用词、词性标注、归一化处理等,并利用 Word2vec 模型生成新语料的词向量空间。在训练 Bi-LSTM-Attention-CRF 模型前,对训练集和测试集进行人工序列标注,采用 3.1.4 小节表 1 示例的方法进行标注,共标注了 23 672 个单词。

4.2 评价指标

通过本文提出的领域新词发现技术可以识别出产品领域新特征,用以领域本体演化中变化的捕捉。因此在对于领域新词发现进行评价时,主要思路是利用本研究采用的 Bi-LSTM-Attention-CRF 模型识别出测试集语料中的智能手机领域新词与人工标注的领域新词进行对比,可以评价出本研究采用模型的优劣。其中,评估的指标为:准确率(Precision)、召回率(Recall)、F 值(F-measure)。公式如下:

$$Precision = \frac{correct_found_new_words}{found_new_words}$$

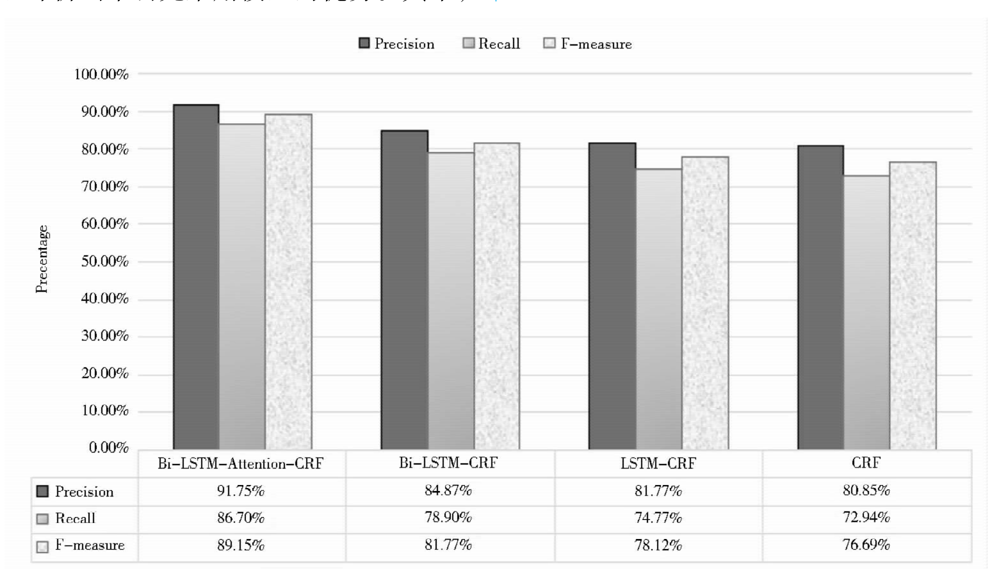
$$Recall = \frac{correct_found_new_words}{correct_new_words}$$

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

其中,correct_found_new_words 表示模型正确识别出领域新词的数量;found_new_words 表示模型识别出领域新词的总数量;correct_new_words 表示新语料中正确的领域新词总数。

4.3 实验结果与讨论

在领域新词发现的实验中,为验证方法的有效性,本研究结合苹果公司官方文档、电商平台产品详情以及领域专家知识,人工标注了领域新词共 654 组,作为语料中正确的领域新词。以 CRF 模型进行过滤新词的方法为 baseline,然后对比 LSTM 结合 CRF 模型、双向 LSTM 结合 CRF 模型,以及本研究所采用的双向 LSTM-Attention-CRF 模型,对比结果如图 7 所示:



	Bi-LSTM-Attention-CRF	Bi-LSTM-CRF	LSTM-CRF	CRF
Precision	91.75%	84.87%	81.77%	80.85%
Recall	86.70%	78.90%	74.77%	72.94%
F-measure	89.15%	81.77%	78.12%	76.69%

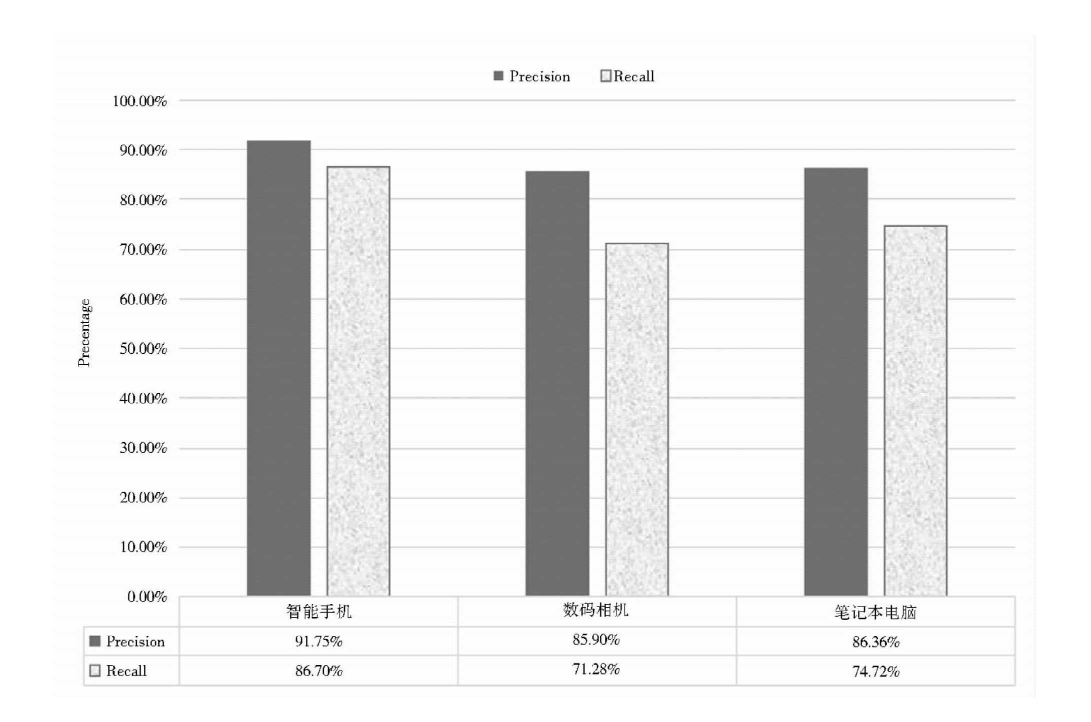
图 7 领域新词发现模型对比结果

由图 7 实验结果可知,利用 Bi-LSTM-Attention-CRF 模型处理得到的领域新词效果最佳,准确率达到 91.75%,F 值达到了 89.15%。

另一方面,为验证本研究采用的领域新词发现在不同数据集的通用性,以类似于智能手机数据集的收集和处理方法,本研究还进行了数码相机产品评论数据集、笔记本电脑产品评论数据集的对比实验。在采用 Bi-LSTM-Attention-CRF 模型的领域新词发现的方法下,不同数据集对于领域新词的识别效果对比见图 8。

由图 8 实验结果可知,本研究采用的领域新词发现的方法在不同数据集的准确率都高于 85%;而由于数码相机产品和笔记本电脑的新特征相对较少,导致召回率相对不高,但都保持在 70% 以上。因此,上述两个实验结果可以验证利用本研究的模型,可以有效地对评论文本中领域新词进行识别和抽取。

在本体演化的实验中,以 iPhone 智能手机领域产



	智能手机	数码相机	笔记本电脑
Precision	91.75%	85.90%	86.36%
Recall	86.70%	71.28%	74.72%

图 8　不同数据集领域新词识别效果对比结果

品本体的演化为例。在前期研究已有领域产品本体的基础上，根据基于领域新词发现的结果，并利用第 3.2 节采用的本体演化的方法，可以对领域本体进行动态调整。经过本体演化的处理后，旧版和新版智能手机领域产品本体在 Protégé 中呈现如图 9、图 10 所示。图 9 是旧版领域产品本体的显示结果；图 10 在 Protégé 中的 OntoGraf 模块中显示的结果，其主要展示出本体演化后领域本体结构的变化，以及本体的类和结构的变化。在图 10 中，其中带有“new”标记且用粗线框标记的矩形框为新版本体新增的类。

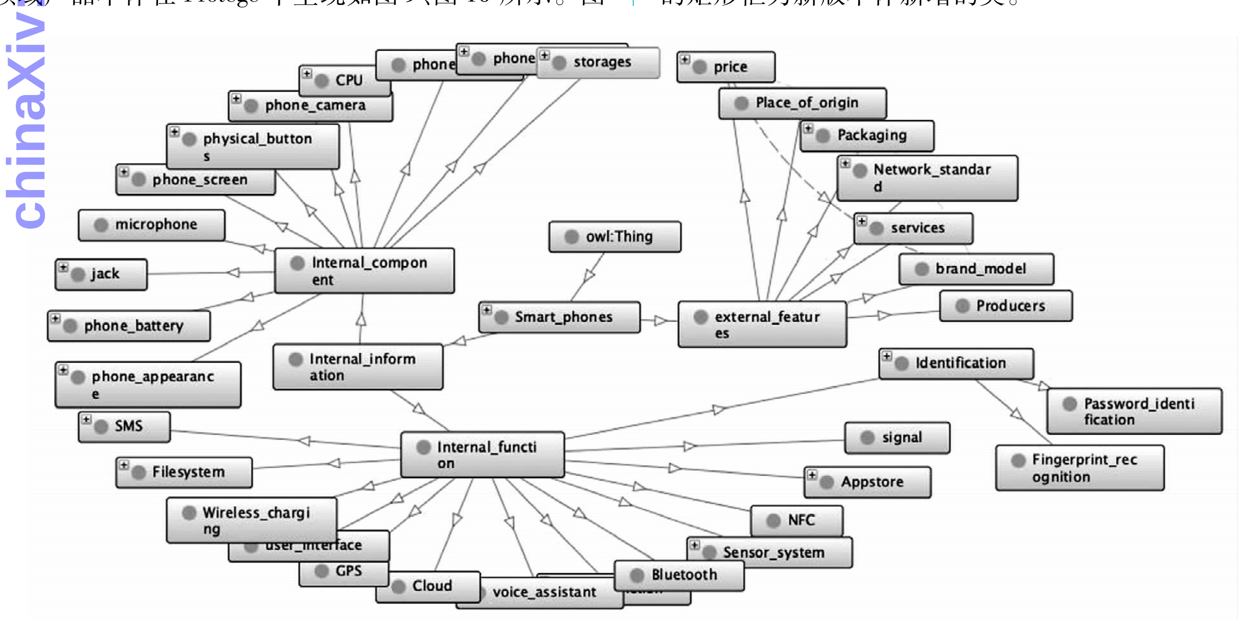


图 9　Protégé 中旧版领域本体结构(部分)示例

由上述新版领域本体结果可知，前期研究主要基于一些早期的用户评论文本数据，由此构建的旧版领域本体会出现应用受限的缺陷，即随着时间和空间的变化，本领域内会出现新特征、新功能，以及要求调整领域本体的结构等需求，旧版领域本体就必须进行调整。新版领域本体是在即时的、最新的用户产品评论

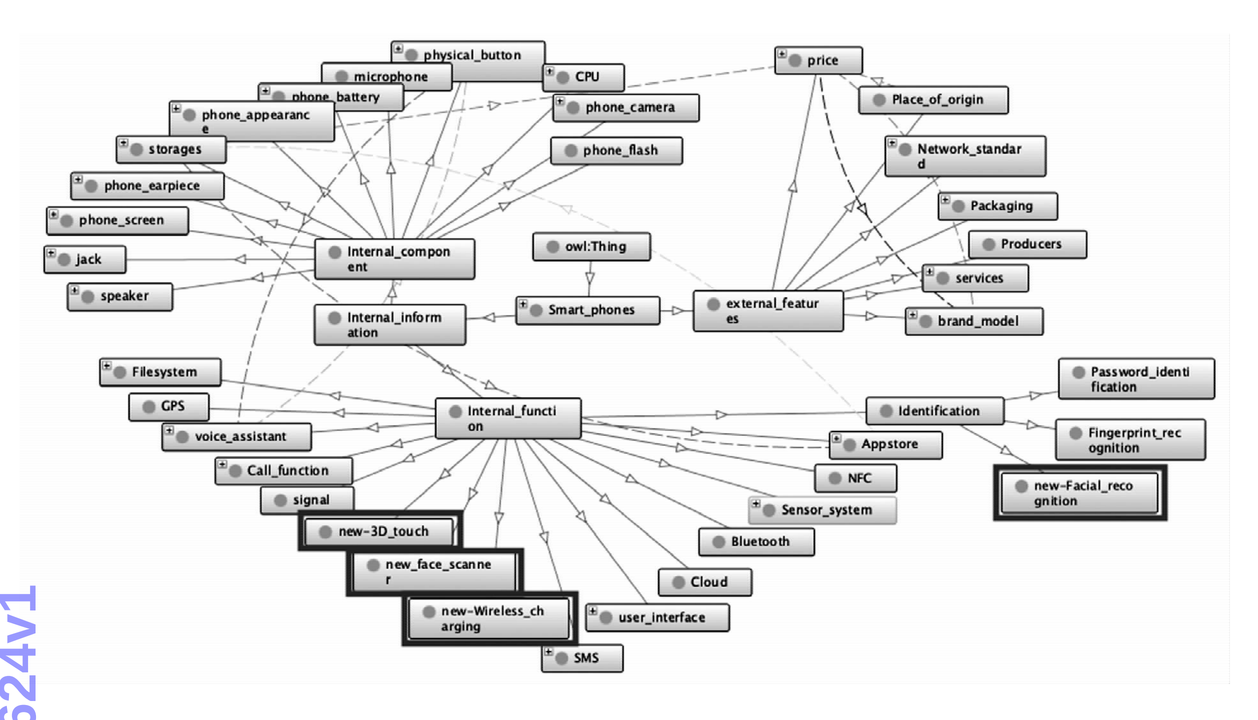


图 10 Protégé 中新版本体中结构(部分)示例

数据背景下演变而来的,在处理新文本时具有可靠性,可以帮助产品设计者根据领域本体中用户关注的产品新特征、新功能、新组件等,优化产品设计,也可以为消费者利用产品评论进行产品购买决策时提供支持,由此以延续并进一步发挥产品评论下领域产品本体的应用价值。

5 结语

随着知识的增加和需求的变化,及时动态调整领域本体对于本体的应用具有重要的意义。本文提出了一种基于领域新词发现的本体演化方法,并以用户产品评论为例论证该方法的有效性。本文从理论意义层面,提出了一个基于领域新词发现的本体演化框架,以及一种融合了 Word2vec 算法、Bi-LSTM-Attention-CRF 算法、K-means 算法的领域新词发现方法,并基于用户产品评论为例验证方法的有效性,利于在非结构化文本数据的知识组织和长期利用。在实践意义层面,本文对于从评论文本中构建的领域产品本体进行了演化,利用领域新词发现技术,从新颖的评论文本语料中挖掘出产品的新特征、新功能、新组件等,为产品设计者优化产品设计提供帮助;同时,也为消费者在利用产品评论进行购买产品时,提供购买决策支持。未来将继续进行不同领域的本体演化的一致性和复用性研究,如融入本体对齐、本体映射等方法。同时将利用新版领域本体进行更多的应用,例如知识推理、知识图谱构建等,充分发挥新版领域本体的价值。

参考文献:

[1] JIN J, LIU Y, JI P, et al. Review on recent advances in information mining from big consumer opinion data for product design[J]. Journal of computing and information science in engineering, 2019, 19(1): 1-19.

[2] 邓斯予,耿骞,靳健,等. 基于产品评论分析的领域知识库构建与应用[J]. 情报理论与实践, 2019, 42(11): 115-122,127.

[3] GENG Q, DENG S, JIA D, et al. Cross-domain ontology construction and alignment from online customer product reviews[J]. Information sciences, 2020, 531:47-67.

[4] CARDOSO S D, SILVEIRA M D, PRUSKI C. Construction and exploitation of an historical knowledge graph to deal with the evolution of ontologies[J]. Knowledge-based systems, 2020, 194(22): 105508.

[5] 陈晶,刘钊,顾进广,等. 本体演化中基于 TFOF 的波及效应分析[J]. 武汉大学学报(理学版), 2020, 66(2):197-204.

[6] BENOMRANE S, SELLAMI Z, AYED M B. An ontologist feedback driven ontology evolution with an adaptive multi-agent system[J]. Advanced engineering informatics, 2016, 30(3): 337-353.

[7] CHEN C, LIU Y, KUMAR M, et al. Energy consumption modelling using deep learning embedded semi-supervised learning[J]. Computers & industrial engineering, 2019, 135:757-765.

[8] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[9] NAGIREDDI V S K, MISHRA S. An ontology based cloud service generic search engine[C]// International conference on computer science & education. Colombo: IEEE, 2013:335－340.
[10] CHEN X, CHEN H, BI X, et al. BioTCM-SE: A semantic search engine for the information retrieval of modern biology and traditional Chinese medicine[J]. Computational and mathematical methods in medicine, 2014,13(2):1－13.
[11] 刘紫玉,杨雨佳,张晓明,等. 基于 DBpedia 的领域本体进化方法研究[J]. 情报杂志, 2017, 36(6):160－166.
[12] 陈晶,刘钊,顾进广,等. 本体演化的波及效应计算优化研究[J]. 计算机应用研究, 2020, 37(8):2366－2370.
[13] 刘毅,王宇,杨德礼. 本体进化驱动的个性化语义搜索研究[J]. 情报学报, 2015, 34(10): 1048－1055.
[14] 刘莹. 基于本体进化和知识检索联动的知识管理系统[J]. 情报科学, 2016, 34(4):62－67.
[15] HUANG C, CAI H, XU L, et al. Data-driven ontology generation and evolution towards intelligent service in manufacturing systems[J]. Future generation computer systems, 2019, 101:197－207.
[16] 刘伟童,刘培玉,刘文锋,等. 基于互信息和邻接熵的新词发现算法[J]. 计算机应用研究, 2019, 36(5):1293－1296.
[17] 郭理,张恒旭,王嘉岐,等. 基于 Trie 树的词语左右熵和互信息新词发现算法[J]. 现代电子技术, 2020, 43(6):65－69.
[18] 王煜,徐建民. 用于网络新闻热点识别的热点新词发现[J/OL]. 计算机应用:1－9[2020－09－12]. http://kns.cnki.net/kcms/detail/51.1307.TP.20200722.1337.002.html.
[19] 杜丽萍,李晓戈,于根,等. 基于互信息改进算法的新词发现对中文分词系统改进[J]. 北京大学学报(自然科学版), 2016, 52(1):35－40.
[20] 周霜霜,徐金安,陈钰枫,等. 融合规则与统计的微博新词发现方法[J]. 计算机应用, 2017, 37(4):1044－1050.
[21] 王馨,王煜,王亮. 基于新词发现的网络新闻热点排名[J]. 图书情报工作, 2015, 59(6):68－74.
[22] 陈梅婕,谢振平,陈晓琪,等. 专利新词发现的双向聚合度特征提取新方法[J]. 计算机应用, 2020, 40(3):631－637.
[23] 张华平,商建云. 面向社会媒体的开放领域新词发现[J]. 中文信息学报, 2017, 31(3):55－61.
[24] 王汀,冀付军,徐天晟. 一种面向中文网络百科非结构化信息的知识获取方法[J]. 图书情报工作, 2016, 60(13):126－133.
[25] 陈先来,韩超鹏,安莹,等. 基于互信息和逻辑回归的新词发现[J]. 数据分析与知识发现, 2019(8):105－113.
[26] 刘昱彤,吴斌,谢韬,等. 基于古汉语语料的新词发现方法[J]. 中文信息学报, 2019, 33(1): 46－55.
[27] 赵志滨,石玉鑫,李斌阳. 基于句法分析与词向量的领域新词发现方法[J]. 计算机科学, 2019, 46(6):29－34.
[28] 黄文明,杨柳青青,任冲. 结合信息量和深度学习的领域新词发现[J]. 计算机工程与设计, 2019, 40(7):1903－1907, 1914.
[29] GREGOR K, DANIHELKA I, GRAVES A, et al. DRAW: a recurrent neural network for image generation[C]//ICML. 15: proceedings of the 32nd international conference on international conference on machine learning. Lille: JMLR, 2015, 37: 1462－1471.
[30] GRAVES A. Supervised sequence labelling with recurrent neural networks[M]// Studies in computational intelligence, SCI 385. Berlin: Springer, 2012:5－13.
[31] PALANGI H, DENG L, SHEN Y, et al. Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval[J]. IEEE/ACM transactions on audio, speech, and language processing, 2015, 24(4):694－707.
[32] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2020－09－16]. https://arxiv.org/pdf/1409.0473.pdf.
[33] 张华丽,康晓东,李博,等. 结合注意力机制的 Bi-LSTM-CRF 中文电子病历命名实体识别[J]. 计算机应用, 2020,40(S1):98－102.
[34] 李纲,潘荣清,毛进,等. 整合 BiLSTM-CRF 网络和词典资源的中文电子病历实体识别[J]. 现代情报, 2020, 40(4):3－12, 58.
[35] MIKOLOV T. Distributed representations of words and phrases and their compositionality[J]. Advances in neural information processing systems, 2013, 26:3111－3119.
[36] 胡甜甜,但雅波,胡杰,等. 基于注意力机制的 Bi-LSTM 结合 CRF 的新闻命名实体识别及其情感分类[J]. 计算机应用, 2020, 40(7):1879－1883.
[37] STOJANOVIC L, MAEDCHE A, MOTIK B, et al. User-driven ontology evolution management[C]// Proceedings of the 13th international conference on knowledge engineering and knowledge management. Ontologies and the semantic Web. Berlin: Springer-Verlag: 2002,285－300.
[38] NOY N F, CHUGH A, LIU W, et al. A framework for ontology evolution in collaborative environments[C]// International semantic web conference. Berlin:Springer, 2006.

作者贡献说明：

耿骞：负责制定论文大纲、论文修改；

邓斯予：负责算法设计与实现、论文撰写；

靳健：负责提出论文思路、论文修改。

图书情报工作
第 65 卷 第 8 期 2021 年 4 月

Integrating Word Semantic Representation and New Word Identification for Domain Ontology Evolution: A Case Study of Product Online Reviews

Geng Qian[1,2] Deng Siyu[2] Jin Jian[2]
[1] Center for Governance Studies, Beijing Normal University, Zhuhai 519087
[2] School of Government, Beijing Normal University, Beijing 100875

Abstract: [Purpose/significance] Due to the inaccuracy and low efficiency in capturing new knowledge and new requirements in traditional ontology evolution, based on domain new word identification, an ontology evolution method is proposed and evaluated by analyzing a large volume of product online reviews. [Method/process] First, a series of natural language processing algorithms were used to pre-process product review text corpus, and the Word2vec algorithm was adopted for word vector embedding. Then, a Bi-LSTM-Attention-CRF algorithm was utilized for the recognition and extraction of new words in a candidate set, and the K-means algorithm was applied for clustering to get the final domain new words. Finally, the Six-Stage evolution process of ontology evolution was invited for analyzing domain ontology evolution. [Result/conclusion] By analyzing smart phone reviews as examples, it can be found that the proposed approach about new word identification presents a higher accuracy and recall rate and a new version of the product ontology in the smart phone domain can be evolved accordingly. It helps designers to optimize feature and function configuration in new product development and consumers to analyze online opinions for purchase decisions.

Keywords: ontology evolution domain new words new word detection attention mechanism Bi-directional Long Short-Term-Memory Conditional Random Field